学校编码：10384　　分类号 _____ 密级 _____
学号：27720121152648　　UDC _____

硕 士 学 位 论 文

基于 Logistic 回归的惩罚性变量选择方法在保险客户识别中的应用

The Application of Penalized Logistic Regression in the Identification of Insurance Clients

张诗悦

指导教师姓名：钟　威　副教授
专 业 名 称：统　计　学
论文提交日期：2015 年　月
论文答辩时间：2015 年　月
学位授予日期：2015 年　月

答辩委员会主席：_____
评　阅　人：_____

2015 年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

摘要

互联网的发展促使各行业产生并储存了大量的数据，如何从大量数据中发掘有用的信息变得越来越重要，变量选择的问题也成为近二十年统计学研究的热点，其在医学和社会科学等领域得到了广泛应用，但在保险领域的应用尚且较少。

目前，我国的保险行业还处于发展初期，消费者对保险产品的认识不足，由于保险营销人员长期以来野蛮的推广方式，使得保险在公众心目中的形象偏离了其给大众带来切身保障的初衷，保险行业的形象亟待重塑，而关键之一就在于对保险客户的准确识别和有效营销。对保险客户的有效识别将节约保险销售的成本，减少对无关客户的干扰，促进保险行业的健康发展。而关于保险客户识别问题的研究较少，因此本文考虑将惩罚性变量选择方法应用于保险客户识别上。

在实际应用中，保险客户数据有其自身的特点，保险客户数据存在不对称性，在大量的客户数据中，只有少部分的客户是目标客户。因此，我们所要研究的是不对称数据结构下的惩罚性变量选择问题。本文模拟产生了不同相关程度的不对称结构数据，在该数据结构下对 LASSO-Logistic 和 SCAD-Logistic 两种方法的变量选择效果和预测准确度进行了比较，结果显示，在不对称数据结构下，SCAD-Logistic 模型和 Lasso-Logistic 模型在变量选择和预测精确度上效果良好，具有适用性。

保险产品还兼具消费性和金融性，影响客户购买保险产品的主要因素与其他产品存在不同，本文将基于 Logistic 回归的变量选择方法应用到保险客户识别上，对影响保险产品购买的主要因素进行研究。本文主要尝试了 LASSO 和 SCAD 这两种惩罚性变量选择方法，对某保险公司提供的客户数据进行了实证分析，结果显示，保险产品的购买受到客户职业、年龄、受教育程度和经济收入的影响，并且，客户在不同险种间存在显著的交叉购买情况，本文的实证研究显示客户购买游艇保险情况的可以显著反映出客户是否购买房车保险。直观上看，房车和游艇均是高端消费，且都用于户外游玩，对客户的经济实力、消费偏好有相似的反映，购买游艇保险也反映了客户较强的保险意识。

关键词：保险客户识别；变量选择；Logistic 模型；

Abstract

The development of Internet brings huge amount of data in many fields. Therefore, to select the useful information from big data has become more and more important. Variable selection is a necessary component of information selection, thus it has been a hotspot of statistical research in recent twenty years. Variable selection has been applied in many fields, such as biology, medicine, social science and finance. However the application in insurance is few.

The process of insurance industry in China is still initial. Since the salesman of insurance take a rude and inefficient way to introduce insurance product to consumers, insurance gives a negative impression to the public, far away from its positive and useful function. With the development of society, insurance industry is eager to rebuild its reputation. One of the key is to change the marketing strategy and to identify the insurance clients accurately. Thus, it can reduce the cost of marketing and alleviate the disturbance of irrelevant clients. The application of variable selection in the identification of insurance clients is quite few. Therefore, this paper will focus on the application of penalized logistic regression in the identification of insurance clients.

In reality, the data of insurance clients has its unique characteristics. There exists one problem that the data of insurance clients is rare. There is only a small part of target clients in the dataset. Therefore, this paper will focus on the application of penalized logistic regression in rare events. This paper mainly uses LASSO and SCAD to do simulation under asymmetric data structures and compares the result of variable selection and prediction. The results show that LASSO and SCAD penalized logistic regression perform well under asymmetric data structures.

The insurance products contain both the features of consumption goods and that of finance product. Therefore, the influence factors of insurance product differ from them. This paper will take advantages of penalized logistic regression into the

research of the influence factors of insurance product. This paper mainly uses LASSO and SCAD to penalize the regression and compare the results of both methods. The results show that the sales of insurance product is influenced by the profession, age, education level and income of the clients. What is more, there exists cross purchase of different insurance product among clients. In our empirical research, we predict the sales of mobile home insurance and find that it is reflected by that of the boat insurance significantly.

Key words: The Identification of Insurance Clients; Variable Selection; Logistic Regression;

目录

Table of Contents

第一章　引言

1.1 研究背景和意义

维克托 迈尔 舍恩伯格和肯尼斯 库克耶在《大数据时代》中提出了大数据的概念，如今大数据在很多领域中已经得到了广泛的应用，如生物、医学、社会科学和环境科学等，在金融保险领域的应用也在逐步开发，已经在预测银行或商业上的诈骗行为、违约行为、客户信用等领域得到了广泛的研究与应用。大数据概念逐步被大家所接受，渗入社会的各个方面。

随着信息化和互联网化，每天都有大量的数据产生，并被记录下来。随着数据的积累，对于数据的运用和发掘成为各个公司竞争的重点，特别是在营销领域。利用客户数据中隐含的信息提高寻找潜在客户的能力，能够为公司提供巨大的优势，这个应用也受到了越来越多的重视，最关注数据分析的领域有航空公司、电子商务平台、银行、保险公司等。公司积累了客户的消费记录和客户的个人信息，如果能从中发掘出客户的偏好，提供匹配的推荐，满足客户的潜在需求，将降低销售成本，使产品推广更有效率。又可以进一步积累更多的客户数据，继续提升产品推广方法，形成一个良好的正循环。

本文重点关注的是数据分析在保险营销领域中的实证研究。随着我国居民收入水平不断提高，居民对于保险的需求逐渐凸显，但是居民对于保险产品的了解还存在不足，由于保险营销人员长期以来野蛮的推广方式，一方面提高了销售成本，另一方面也使得保险行业在公众心目中的形象偏离了保险给大众带来切身保障的初衷。保险市场是一个高速发展且竞争激烈的市场，保险以大数定律为原则，需要吸引尽可能多的客户，只有以多样化的风险为经营对象，扩大规模，才能进一步发展，因而如何提高销售效率是保险行业积极关注并急需解决的问题。

保险公司关注的是找出可以用于预测客户险种偏好的信息，发掘潜在的客户。目前，我国的保险行业发展还不够健全，保险推广的渠道还比较传统，但在互联网金融的带动下，已经逐步向电商发展，也产生了更多的数据，海量数据带来了大量信息的同时，也包含了很多无效的信息，如何选取其中有用的信息成为难点所在。在大数据时代，数据的维数很高，但大部分的解释变量都和被解释变量无

关，如何从大量数据中发掘有用的信息变得越来越重要，变量选择的问题应运而生，成为近二十年统计学研究的热点。

1.2 本文的主要工作

本文针对保险行业的特点，对基于 Logistic 回归的惩罚性变量选择方法在保险营销中的应用进行了探讨，在实际中存在一个难点，即保险客户数据结构存在严重的不对称性，在我们能够获得的海量客户数据中，只有少部分客户是我们的目标客户，如何运用不对称数据进行变量选择和预测是变量选择问题在保险客户识别中应用的特别之处，与其他领域的应用存在不同之处。本文在研究中考虑了这个问题，对不对称数据结构下带有惩罚项的 Logistic 回归进行了模拟，结果显示，在不对称数据结构下，SCAD-Logistic 模型和 Lasso-Logistic 模型在变量选择和预测精确度上效果良好，具有适用性。

本文基于一组真实的保险公司的客户数据，对保险行业的客户识别和险种交叉销售进行了实践探索，应用了基于 Logistic 回归的惩罚性变量选择方法，对客户购买房车保险的情况进行分析。比较了 LASSO 和 SCAD 这两种变量选择方法的应用效果，对于筛选出的变量的实际含义进行分析。保险产品是一个兼具消费性和金融性的产品，与其他产品的销售有所区别。客户在保险购买行为中，反应出客户的经济状况和保险意识等，很可能存在对两种或几种保险产品均存在需求的情况，如何通过客户之前的保险购买行为来预测客户的需求，这是保险营销中一个值得关注的话题。本文通过实证方法，对客户是否会购买房车险的预测结果进行评估，找出对保险产品销售影响显著的因素，结果显示，保险产品的购买受到客户职业、年龄、受教育程度、家庭经济收入的影响，同时，客户在不同险种间存在显著的交叉购买情况，客户购买房车保险的情况受到购买游艇保险的情况影响显著。基于该实证研究，对保险产品的营销提出现实合理的意见。同时，本文还对带有惩罚项的 Logistic 回归的预测准确度和稳健性进行分析，说明了该方法在实证研究中的适用性。

1.3 本文的结构

本文主要包括以下六个部分。第一部分是引言，介绍了本文的选题背景和意